AUTOMATIC SUMMARY EVALUATION. ROUGE MODIFICATIONS

Liana Ermakova
Perm State National Research University
e-mail: Liana87@mail.ru

Abstract

Nowadays there is no common approach to summary. Manual evaluation is expensive and subjective and it is not applicable in real time or on a large corpus. Widely used approaches involve little human efforts and assume comparison with a set of reference summaries. We tried to overcome drawbacks of existing metrics such as ignoring redundant information, synonyms and sentence ordering. Our method combines edit distance, ROUGE-SU and trigrams similarity measure enriched by weights for different parts of speech and synonyms. Since nouns provide the most valuable information, each sentence is mapped into a set of nouns. If the normalized intersection of any pair is greater than a predefined threshold the sentences are penalized. Doing extracts there is no need to analyze sentence structure but sentence ordering is crucial. Sometimes it is impossible to compare sentence order with a gold standard. Therefore similarity between adjacent sentences may be used as a measure of text coherence. Chronological constraint violation should be penalized. Relevance score and readability assessment may be combined in the F-measure. In order to choose the best parameter values machine learning can be applied.

Keywords: *Automatic summary evaluation, ROUGE, summarization, edit distance, readability, sentence ordering, redundant information.*

1. INTRODUCTION

Automatic summary evaluation is an important but not solved problem. One of the reasons is the fact that the key concepts such as relevance, user information needs etc. are not well defined. Information importance may be estimated in several ways: conventional significance, relevance to a query or how well this information satisfies user needs [1]. An information retrieval system may be assessed qualitatively or quantitatively [2]. Usually qualitative evaluation is related to expert assessment while quantitative means automatic. Expert assessment is expensive and subjective. Often judges disagree [3]. Manual evaluation on a large scale collection is impossible. It is not applicable in real time (e.g. for algorithm tuning). Automatic methods without any human intervention are not used since these techniques provide low results. Therefore the most practical approaches involve little human efforts. Commonly used metrics are based on the comparison of candidate summaries with reference ones. ROUGE metrics are the most popular. Nevertheless they suffer from a number of drawbacks. ROUGE metrics cannot deal with synonyms. The major shortcoming of ROUGE is that they do not consider redundant information. Therefore our goal is to

improve ROUGE such wise it can treat different order of words and sentence as well as redundant information.

The paper is organized as follows. The next section describes existing evaluation methods including human assessment and comparison with reference summaries. Special attention is paid to ROUGE metrics. After that we will present approaches to readability evaluation. The fourth section provides our ROUGE modifications.

2. OVERVIEW OF SUMMARY EVALUATION METHODS

2.1. Human Evaluation

Human judgment includes assessment of readability, coherence, conciseness, content, grammar, recall, pithiness etc. [4][5][6]. Often these parameters are not numerically expressed but summaries are ranked according to them [3]. The retained information may be evaluated in the following way: one assessor team develop a set of questions based on the input texts, another team should answer these question reading only summaries [7]. An assessor may be asked to evaluate the importance of each sentence/passage ("usefulness assessment"). This annotation allows to generate summaries with predefined compression rate, expert extracts which may be used as a gold standard [4].

One of the significant drawbacks of human assessment is that judgment may be quite different. Normally assessment agreement is 70% due to the fact that judges may have different opinions about summary quality and evaluation metrics [3]. Cohen's kappa is a statistical measure of inter-rater for qualitative [8]. Kappa equals to one means full agreement. If kappa equals to zero, the agreement is coincidence [4]:

$$K = \frac{P(A) - P(E)}{1 - P(E)} \quad (1)$$

where $P(A)$ is agreement rate, $P(E)$ is expected chance agreement rate [8][4].

2.2. Comparison with Reference Summaries

Reference summaries allow to compute the metrics commonly used in IR recall and precision:

$$Recall = \frac{Correct}{Correct + Missed'} \quad (2)$$

$$Precision = \frac{Correct}{Correct + FalsePositive'} \quad (3)$$

where *Correct* is the number of sentences appearing in both reference and candidate summaries, *Missed* is the number of sentences presented in the

reference summary but missed in the candidate summary, *FalsePositive* — is the number of sentences presented in the candidate summary but missed in the reference summary [3]. Recall and precision may be integrated into the F-measure [5]:

$$F = \frac{(\beta^2 + 1) \times Recall \times Precision}{\beta^2 \times Precision + Recall} \quad (4)$$

$$\beta^2 = \frac{1 - \alpha}{\alpha}, \alpha \in [0, 1] \quad (5)$$

F-measure is widely used in IR however it is less useful in summary evaluation since it depends on the recall. Search engine result is potentially infinite while a summary is limited. Moreover this measure cannot be applied to abstract assessment since abstracts suppose reformulation of original sentences.

Similarity may be estimated as cosine, dice or Jaccard coefficient, as well as the number of shared n-grams or longest common subsequence [6].

The metric BLEU commonly used for machine translation evaluation is also suitable for assessment of any generated text. As ROUGE, BLEU is also estimated as the number of shared n-grams [5].

In order to evaluate the impact of algorithm tuning a system is compared with a baseline which do not include the tuning [4].

At the forum for evaluation of information retrieval INEX 2011 summaries are compared with the pool of relevant passages provided by humans. As the distance Kullback-Leibler divergence or simple log difference are used:

$$\sum \log \left(\frac{\max\left(P(t \mid reference), P\left(t \mid summary\right)\right)}{\min\left(P(t \mid reference), P\left(t \mid summary\right)\right)} \right), \quad (6)$$

where t is a term of a document, *reference* is a pool of relevant passages, *summary* is a candidate summary.

S.Tratz and E.Hovy proposed to use Basic Elements (BEs) which mean almost the same thing but are expressed differently. BEs are able to deal with paraphrasing [9]. A BE is a syntactic unit up to 3 words with associated tags such as NER and POS. BEs can take into account lemmas, synonyms, hyponyms and hyperonyms, identical prepositional phrases, spelling variants, nominalization and denominalization (derivation in WordNet), transformations like prenominal noun-prepositional phrase, noun swapping for IS-A type rules, pronoun transformations, pertainym adjective transformation [9].

Meteor evaluation metric is also able to treat spelling variants, WordNet synsets and paraphrase tables [10]. Meteor distinguishes function and

content words. However, this system is designed for machine translation evaluation and fails to deal with texts of different length.

In practice resampling methods are often used, e.g. jackknifing (using subsets of available data) or bootstrapping (random replacement of points in the data set). In this case assessment is the mean of all computed values [9].

Summaries may be evaluated according to compression rate (CR) or retention rate (RR):

$$CR = \frac{Length(S)}{Length(T)} \tag{7}$$

$$RR = \frac{Info(S)}{Info(T)}, \tag{8}$$

where S is a candidate summary and T is an original text. A good summary should have low CR and high RR [3]. CR is well defined and can be easily computed while RR estimation is more problematic since it involves less formalized concepts.

In order to evaluate the quality of assessment metrics one can apply the correlation between expert results and candidate metrics (e.g. Kendall, Spearman or Pearson coefficients). A good metric should give low score to summaries which have low score according to human judgment and high score otherwise [5].

2.3. ROUGE Metrics

One of the most efficient metrics of summary evaluation is ROUGE (Recall-Oriented Understudy for Gisting Evaluation). ROUGE is also based on comparison with reference summaries [5]. ROUGE was proposed by Chin-Yew Lin in 2004 to evaluate summaries but it is also used to evaluate the quality of machine translation. Let look at some ROUGE metrics.

ROUGE-N shows the n-grams recall [5]:

$$ROUGE\text{-}N = \frac{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in ReferenceSummaries} \sum_{gram_n \in S} Count(gram_n)}, \tag{9}$$

where $Count_{match}(gram_n)$ is the maximum number of shared n-grams in the set of reference summaries r_i and in the candidate one S, and $Count(gram_n)$ is the number of n-grams in the set of references. ROUGE-N implies that a summary get higher score as it contains more n-grams co-occurring with reference summaries. ROUGE-N_{multi} compute pairwise n-gram recall and takes the maximal value [5]:

$$ROUGE\text{-}N_{multi} = \text{argmax}_i ROUGE\text{-}N(S, r_i) \tag{10}$$

Another method ROUGE-L is based on the searching of the longest common substring (LCS) shared by two sentences [5]:

$$ROUGE\text{-}L = F_{lcs} = \frac{(\beta^2+1)\times R_{lcs}\times P_{lcs}}{\beta^2\times P_{lcs}+R_{lcs}} \quad (11)$$

$$R_{lcs} = \frac{LCS(X,Y)}{m} \quad (12)$$

$$P_{lcs} = \frac{LCS(X,Y)}{n}, \quad (13)$$

where $LCS(X,Y)$ is the longest common substring of the sentences X and Y, m and n are the lengths of X and Y respectively, and $\beta=P_{lcs}/R_{lcs}$. *ROUGE-L*=1 if $X=Y$, if there is no shared subsequence *ROUGE-L*=0. *ROUGE-L*includes the longest common n-gram and there is no need to compute its length in advance. *ROUGE-L* allows to compare the sentence structure but only with respect to the longest shared part. For the whole texts *ROUGE-L* can be estimated by the formulas [5]:

$$ROUGE\text{-}L = F_{lcs} = \frac{(\beta^2+1)\times R_{lcs}\times P_{lcs}}{\beta^2\times P_{lcs}+R_{lcs}} \quad (14)$$

$$R_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\cup}(X,Y)}{m} \quad (15)$$

$$P_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\cup}(X,Y)}{n}, \quad (16)$$

where u is the number of sentences in the reference summary, v is the number of sentences in the candidate summary, m is the total number of words in the reference summary, n is the total number of words in the candidate summary, $LCS_{\cup}(r_i,C)$ is the LCS score of the union longest common subsequence between reference sentence r_i and candidate summary C. For example, let $r_i=w_1w_2w_3w_4w_5$, and $C=c_1c_2$, $c_1=w_1w_2w_6w_7w_8$, $c_2=w_1w_3w_8w_9w_5$, then LCS for r_i,c_1 is w_1w_2, and for r_i,c_2 LCS is $w_1w_3w_5$. The union is $w_1w_2w_3w_5$. $LCS_{\cup}(r_i,C)=4/5$ [5].

Normalized pairwise comparison $LCS_{MEAD}(S_1,S_2)$ [11] is similar to *ROUGE-L* when $\beta=1$, but LCS_{MEAD} takes the maximal value of *LCS*, while *ROUGE-L* deals with the union of *LCSs* [9].

$$LCS(S_1,S_2) = \\ = \frac{\sum_{s_i\in S_1}\max_{s_j\in S_2}LCS(s_i,s_j)+\sum_{s_j\in S_2}\max_{s_i\in S_1}LCS(s_i,s_j)}{\sum_{s_i\in S_1}length(s_i)+\sum_{s_j\in S_2}length(s_j)} = \\ = \frac{2\times\sum_{s_i\in S_1}\max_{s_j\in S_2}LCS(s_i,s_j)}{m+n} \quad (17)$$

$$LCS_{MEAD}(S_1, S_2) = F_{lcs} = \frac{(\beta^2 + 1) \times R_{lcs\_MEAD} \times P_{lcs\_MEAD}}{\beta^2 \times P_{lcs\_MEAD} + R_{lcs\_MEAD}} \quad (18)$$

$$R_{lcs\_MEAD} = \frac{\sum_{s_i \in S_1} \max_{s_j \in S_2} LCS(s_i, s_j)}{m} \quad (19)$$

$$P_{lcs\_MEAD} = \frac{\sum_{s_j \in S_2} \max_{s_i \in S_1} LCS(s_i, s_j)}{n} \quad (20)$$

One of the serious shortcomings of *LCS* is the fact that it does not consider the distance between words. Weighted *LCS* — *WLCS* — takes into account the length of consecutive matches [5]:

The weighting function *f* should satisfy the following constraint:

$$(\forall x, y \in N): f(x + y) > f(x) + f(y)$$

That is to say consecutive matches should have higher score than non-consecutive ones. *f* may be the linear $f(k)=\alpha k-\beta$, $\alpha>0$, $\beta>0$, polynomial or quadratic function $f(k)=k^2$. In this case F-measure is estimated as follows [5]:

$$F_{wlcs} = \frac{(\beta^2 + 1) \times R_{wlcs} \times P_{wlcs}}{\beta^2 \times P_{wlcs} + R_{wlcs}} \quad (21)$$

$$R_{wlcs} = f^{-1}\left(\frac{WLCS(X, Y)}{f(m)}\right) \quad (22)$$

$$P_{wlcs} = f^{-1}\left(\frac{WLCS(X, Y)}{f(n)}\right) \quad (23)$$

LCS based algorithms are a special case of edit distance [12].

The metric ROUGE-S is based on the counting of shared bigrams the elements of which may be separated by arbitrary number of other words. To compute ROUGE-S the formulas 24-26 are applied:

$$F_{skip2} = \frac{(\beta^2 + 1) \times R_{skip2} \times P_{skip2}}{\beta^2 \times P_{skip2} + R_{skip2}} \quad (24)$$

$$R_{skip2} = \frac{SKIP2(X, Y)}{C(m, 2)} \quad (25)$$

$$P_{skip2} = \frac{SKIP2(X, Y)}{C(n, 2)} \quad (26)$$

where $C(n,k)$ is the binomial coefficient $\binom{n}{k}$, and $SKIP2(X,Y)$ is the number of common bigrams with arbitrary distance in the texts *X* and *Y* respectively. The distance may be limited by d_{skip}. Sometimes unigram smoothing is applied (ROUGE-SU) [5].

For the Russian language a special metric ROUGE-RUS was developed. It considers Russian morphology as well as synonyms (thesaurus). A

summary is compared with a set of references and the result assessment is the mean value of all obtained scores [6].

3. READABILITY EVALUATION

Traditional methods of readability evaluation are based on familiarity of terms and syntax complexity [13]. Word complexity mays be estimated by humans [14][15][16] or according to its length [17]. Researches also propose to use language models [13][18].

Usually assessors assign score to the readability of text in some range [19].

Syntactical errors, unresolved anaphora, redundant information and coherence influence readability and therefore the score may depends on the number of these mistakes [20].

Different non-parametric rank correlation coefficients (e.g. Kendall, Spearman or Pearson coefficients) may be used to find the dependence [21]. However as it is shown in [22] Kendall coefficient is the most suitable for sentence ordering assessment:

$$\tau = 2\frac{number(agreement) - number(disagreement)}{N(N-1)} \qquad (27)$$

Since sentence ordering may be "correct", but not necessary unique, it is advisable to consider an average value of different results [22].

BLEU and edit distance may be applied for relevance judgment as well as for readability evaluation. Another set of methods is based on syntax analysis [23][24][25]. Syntactical methods may be combined with statistics (e.g. sentence length, the depth of a parse tree, omission of personal verb, rate of prepositional phrases, noun and verb groups etc.) [26]. Last methods are suitable only for the readability evaluation of a particular sentence and therefore they cannot be used for extracts assessment.

4. PROPOSED ROUGE MODIFICATIONS

Sequential comparison of a candidate summary with a gold standard is not robust with respect to the number of sentences and their order. Searching for shared n-grams in the entire summary may cause the assignment of the highest score to the meaningless bag of words if n is small. In case of pairwise comparison the highest score may be assigned to the summary which consists of single sentence repeating several times. In order to avoid this effect one can use the alignment of sentences similar as in machine translation.

We believe that the most efficient metric is WLCS (this agrees with evaluation of metrics performed in 2004 [5]). However it has

serious drawbacks. Let look at them more deeply. WLCS is not robust with respect to the word order but in real text the almost same idea may be expressed in several ways (e.g. "Mickey Mouse was created by Walt Disney" and "Walt Disney created Mickey Mouse"). Searching for WLCS in the entire summary may assign different score to the summaries with different sentence order. However, as it was already mentioned the correct sentence order may not be unique [22]. Partially the problem may be solved by unigram smoothing [5]. WLCS consider the length of consequent matches but it does not take into account the distance between them. WLCS assigns the same score to the sequences ABCQWERTYDEF and ABCQDEF if the gold standard is ABCDEF. Therefore we propose to combine edit distance and WLCS by applying the increasing penalty function for the distance between matching elements. Edit distance is the minimum number of operations (insertion, deletion, substitution, or a transposition of two adjacent symbols) needed to transform one string into the other. The Wagner-Fischer algorithm computes edit distance in $O(nm)$ time, where n and m are the length of the strings [27]. WLCS consider only the longest substring and therefore it is advisable to combine it with another method, e.g. ROUGE-S smoothed by unigrams (ROUGE-SU). We propose to modify ROUGE-SU by introducing the increasing penalty function for the distance between elements. ROUGE-3 seems to be useful to readability evaluation. Thus, our metrics is the weighted sum of modified edit distance, ROUGE-SU and ROUGE-3.

ROUGE does not take into account the difference between exact token matching, shared lemmas and synonyms. Thereby we added the coefficients for (in decreasing order): (1) exact matching; (2) matching of lemmas; (3) matching of stems; (4) synonyms; (5) substitution by a hyperonym (only for nouns); (6) substitution by a hyponym (only for nouns). We adopted the idea of H. G. Silber and K. F. Mccoy that nouns provide the most valuable information [29] and that is why we propose to introduce coefficients to distinguish the impact of nouns, other significant words and stop-words. Another modification involves anaphora resolution. For example, it may be performed by Stanford parser. For each sentence the mention (contextual synonym) giving the best score is chosen.

The major disadvantage of ROUGE metrics seems to be the ignoring of redundant information. Each sentence should be mapped into a set of nouns. These sets are compared pairwise and if the normalized intersection is greater than a predefined threshold the sentences are penalized.

Let's consider the following examples. Example 1 is a reference summary and Example 2, Example 3, Example 4 and Example 5 are candidate summaries.

Example 1

Skyfall is the twenty-third spy film in the James Bond series, produced by Eon Productions for MGM, Columbia Pictures and Sony Pictures Entertainment.
Directed by Sam Mendes, it features Daniel Craig's third performance as James Bond and Javier Bardem as Raoul Silva, the film's villain.
Skyfall will also be the first James Bond film to be released in IMAX venues.

Example 2

Skyfall will also be the first James Bond film to be released in IMAX venues.
Skyfall is the twenty-third spy film in the James Bond series, produced by Eon Productions for MGM, Columbia Pictures and Sony Pictures Entertainment.
The film's release will coincide with the 50th anniversary of the Bond film series, which began with Dr. No in 1962.
Directed by Sam Mendes, it features Daniel Craig's third performance as James Bond and Javier Bardem as Raoul Silva, the film's villain.

Example 3

film to James Bond also be Skyfall will the IMAX venues first be released in .
Skyfall is the twenty-third spy film in the James Bond series, produced by Eon Productions for MGM, Columbia Pictures and Sony Pictures Entertainment.
Directed by film's villain Sam Mendes, it features Javier Bardem third performance and as Raoul Silva, the Daniel Craig's as James Bond.

Example 4

Skyfall will also be the first James Bond film to be released in IMAX venues.
Skyfall is the twenty-third spy film in the James Bond series, produced by Eon Productions for MGM, Columbia Pictures and Sony Pictures Entertainment.
Directed by Sam Mendes, it features Daniel Craig's third performance as James Bond and Javier Bardem as Raoul Silva, the film's villain.
Skyfall will also be the first James Bond film to be released in IMAX venues.

Example 5

Skyfall will be the first James Bond film to be released in IMAX venues.
It is produced by Eon Productions for MGM, Columbia Pictures and Sony Pictures Entertainment.
Skyfall is the twenty-third spy film in the series.
Skyfall is directed by Sam Mendes.
Skyfall features Daniel Craig's third performance as James Bond.
Javier Bardem plays the film's villain Raoul Silva.

Meteor 1-3 failed to treat these examples since they have different length [10]. ROUGE package showed results presented in Table 1.

Table 1.

	Example 2	**Example 3**	**Example 4**	**Example 5**
ROUGE-1	R:1.00000 P:0.74118 F:0.85135	**R:1.00000** **P:1.00000** **F:1.00000**	R:1.00000 P:0.80769 F:0.89362	R:0.90476 P:0.89062 F:0.89763
ROUGE-2	R:0.96774 P:0.71429 F:0.82192	R:0.74194 P:0.74194 F:0.74194	**R:1.00000** **P:0.80519** **F:0.89208**	R:0.72581 P:0.71429 F:0.72000
ROUGE-3	R:0.93443 P:0.68675 F:0.79167	R:0.52459 P:0.52459 F:0.52459	**R:1.00000** **P:0.80263** **F:0.89051**	R:0.57377 P:0.56452 F:0.56911
ROUGE-W-1.2	R:0.45255 P:0.61980 F:0.52313	R:0.32251 P:0.59594 F:0.41852	**R:0.45255** **P:0.67542** **F:0.54197**	R:0.35158 P:0.63951 F:0.45372
ROUGE-SU*	R:0.72804 P:0.40148 F:0.51755	R:0.63722 P:0.63722 F:0.63722	**R:1.00000** **P:0.65422** **F:0.79097**	R:0.57419 P:0.55652 F:0.56522
ROUGE-L	R:1.00000 P:0.74118 F:0.85135	R:0.77778 P:0.77778 F:0.77778	**R:1.00000** **P:0.80769** **F:0.89362**	R:0.85714 P:0.84375 F:0.85039

As it is given in Table 1 ROUGE-1 assigned the maximal score to Example 3. But the summary is quite poor since it is unreadable and contains many syntactical errors. According other metrics the best summary is Example 4 which has a redundant sentence. However we can see that Example 2 may be even better because it includes additional information. Example 5 differs from the reference summary only by paraphrases, but it has lower weight than the worse summary Example 3 according to ROUGE-SU*, ROUGE-2 and ROUGE-1. Penalty of Example 4 for the redundant sentence decrease the score by $length_{sentence}/length_{summary}*weight_{redundancy}$=15/75*1=0.2.

The final score assign by the modified metric is presented in Table 2.

Table 2.

	Example 2	**Example 3**	**Example 4**	**Example 5**
ROUGE_MOD	0.23	0.12	0.22	0.27

ROUGE evaluates only the informative content of a summary. However, R. Barzilay, N. Elhadad, and K. R. McKeown showed that sentence order influences a lot on text perception [19]. In case of comparison with a pool of reference sentences (not individual summaries) it is impossible to use rank correlation coefficient since there is no order. Moreover, since correct sentence order may be not unique, providing all possible correct orders may be expensive. Our hypothesis is that a good sentence order supposes similarity between adjacent sentences. However it should not violate chronological constraint. Many modern parsers (e.g. Stanford Core NLP) include a component normalizing dates. Thus, we propose to estimate the quality of sentence order as the sum of distances between adjacent sentences (e.g. cosine coefficient). Chronological constraint violation should be penalized. ROUGE-L and ROUGE-1 are not sensible to sentence order in contrast to ROUGE-SU*, ROUGE-W-1.2, ROUGE-2 and ROUGE-3.

It seems to be useful to combine informational content score and readability assessment, e.g. by using F-measure:

$$F = \frac{Relevance \times Readability}{\alpha \times Relevance + (1 - \alpha) \times Readability}$$

The values of the parameters may be estimated by machine learning techniques on a corpus of the evaluated summaries.

5. CONCLUSION

Nowadays there is no common approach to summary evaluation though there are several metrics of quality assessment. Some techniques involve human intervention. Manual evaluation is expensive and subjective and it is not applicable in real time or on a large corpus. Widely used approaches involve little human efforts and assume comparison with a set of reference summaries. ROUGE is one of those. However, it has several drawbacks such as ignoring redundant information, synonyms and sentence ordering. Thereby we propose the method of summary evaluation which combines edit distance, ROUGE-SU and trigrams similarity measure enriched by weights for different parts of speech and synonyms. Since nouns provide the most valuable information [29], each sentence is mapped into the set of nouns. If the normalized intersection of any pair is greater than a predefined threshold the sentences are penalized. Doing extracts there is no need to analyze sentence structure but sentence ordering is crucial. Sometimes it is impossible to compare sentence order with a gold standard. Therefore similarity between adjacent sentences may be used as a measure of text coherence. Chronological constraint violation should be penalized. Relevance score and readability

assessment may be combined in the F-measure. In order to choose the best parameter values machine learning can be applied.

REFERENCES

1. **K. Filippova**, "Text-to-Text Generation," RuSSIR/EDBT 2011, 2011.
2. **K. S. Jones and J. R. Galliers**, Evaluating Natural Language Processing Systems: An Analysis and Review, vol. 24. Springer-Verlag New York, Inc. Secaucus, 1996.
3. **S. Gholamrezazadeh, M. A. Salehi, and B. Gholamzadeh**, "A Comprehensive Survey on Text Summarization Systems," Computer Science and its Applications, pp. 1–6, 2009.
4. **H. Saggion, D. Radev, S. Teufel, W. Lam, and S. M. Strassel**, "Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Cross-lingual Environment," LREC, pp. 747–754, 2002.
5. **C.-Y. Lin**, "ROUGE: A Package for Automatic Evaluation of Summaries," Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pp. 74–81, 2004.
6. **С.Д. Тарасов**, "Метод тематического связанного ранжирования для задач автоматического сводного реферирования научно-технических информационных сообщений," Балтийский Государственный Технический Университет «ВОЕНМЕХ» им. Д.Ф.Устинова, Санкт-Петербург, 2011.
7. **Y. Seki**, "Automatic Summarization Focusing on Document Genre and Text Structure," ACM SIGIR Forum, vol. 39, no. 1, pp. 65–67, 2005.
8. **J. Carletta**, "Assessing agreement on classification tasks: The kappa statistic," Computational Linguistics, vol. 22, pp. 249–254, 1996.
9. **E. Hovy and S. Tratz**, "Summarization Evaluation Using Transformed Basic Elements," Proceedings TAC 2008, 2008.
10. **M. Denkowski and A. Lavie**, "Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems," Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation, pp. 85–91, 2011.
11. **D. Radev, S. Teufel, H. Saggion, W. Lam, J. Blitzer, A. Elebi, H. Qi, E. Drabek, and D. Liu**, "Evaluation of text summarization in a cross-lingual information retrieval framework," Center for Language and Speech Processing, Johns Hopkins University, Baltimore, 2002.
12. **S. Bangalore, O. Rambow, and S. Whittaker**, "Evaluation metrics for generation," Proceedings of the first international conference on, pp. 1–8, 2000.
13. **K. Collins-Thompson and J. Callan**, "A Language Modeling Approach to Predicting Reading Difficulty," Proceedings of HLT/NAACL, vol. 4, 2004.
14. **A. J. Stenner, I. Horablin, D. R. Smith, and M. Smith**, "The Lexile Framework. Durham, NC: Metametrics," 1988.
15. **J. S. Chall and E. Dale**, Readability revisited: The new Dale-Chall readability. Cambridge: MA: Brookline Books, 1995.
16. **E. Fry**, "A readability formula for short passages," Journal of Reading, vol. 8, pp. 594–597, 1990.
17. **J. Tavernier and P. Bellot**, "Combining relevance and readability for INEX 2011 Question-Answering track," pp. 185–195, 2011.

18. **L. Si and J. Callan**, "A statistical model for scientific readability," Proceedings of the tenth international conference on Information and knowledge management, pp. 574–576, 2001.
19. **R. Barzilay, N. Elhadad, and K. R. McKeown**, "Inferring Strategies for Sentence Ordering in Multidocument News Summarization," Journal of Artificial Intelligence Research, no. 17, pp. 35–55, 2002.
20. **E. SanJuan, V. Moriceau, X. Tannier, P. Bellot, and J. Mothe**, "Overview of the INEX 2011 Question Answering Track (QA@INEX)," Focused Retrieval of Content and Structure, 10th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2011), Geva, S., Kamps, J., Schenkel, R. (Eds.), 2012.
21. **G. Lebanon and J. Lafferty**, "Cranking: Combining rankings using conditional probability models on permutations," Machine Learning: Proceedings of the Nineteenth International Conference, pp. 363–370, 2002.
22. **M. Lapata**, "Probabilistic Text Structuring: Experiments with Sentence Ordering," Proceedings of ACL, pp. 542–552, 2003.
23. **A. Mutton, M. Dras, S. Wan, and R. Dale**, "Gleu: Automatic evaluation of sentence-level fluency," ACL'07, pp. 344–351, 2007.
24. **S. Wan, R. Dale, and M. Dras**, "Searching for grammaticality: Propagating dependencies in the viterbi algorithm," Proceedings of the Tenth European Workshop on Natural Language Generation, 2005.
25. **S. Zwarts and M. Dras**, "Choosing the right translation: A syntactically informed classification approach," Proceedings of the 22nd International Conference on Computational Linguistics, pp. 1153–1160, 2008.
26. **J. Chae and A. Nenkova**, "Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text," Proceedings of the 12th Conference of the European Chapter of the ACL, pp. 139–147, 2009.
27. **R. A. Wagner and M. J. Fischer**, "The string-to-string correction problem," Journal of the ACM, vol. 21, no. 1, p. 168—173, 1974.
28. **H. G. Silber and K. F. Mccoy**, "Efficiently computed lexical chains as an intermediate representation for automatic text summarization," Computational Linguistics - Summarization, vol. 28, no. 4, pp. 1–11, 2002.